DOCUMENT RESUME

ED 419 012 TM 028 312

AUTHOR Camara, Wayne J.
TITLE Constraints and Limitations in Evaluating Math Curricular Reform Efforts: Pacesetter Math Case Study.
PUB DATE 1998-04-00
NOTE 23p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Diego, CA, April 13-17, 1998).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Case Studies; *Curriculum Development; *Educational Change; *Evaluation Methods; Evaluation Problems; *High School Students; High Schools; Honors Curriculum; *Mathematics Curriculum; Tables (Data)
IDENTIFIERS *College Entrance Examination Board; *Constraints; Reform Efforts

ABSTRACT

This paper addresses the challenges and strategies of evaluating curricular reforms in secondary schools by presenting a case study of the College Board's Pacesetter Math course, a fourth level course that was entering its third year in 1995-96. The Pacesetter math course is intended to be an alternative to more traditional pre-calculus courses, and is designed for a range of students with differing interests, career intentions, and mathematics preparation. The culminating assessment is completed over two course periods, and is a standard part of the course. New Pacesetter teachers complete an intensive staff training course. Pacesetter math, at the time of the evaluation, was being implemented in 46 school districts and 130 schools. With no available control group and no pre/post design, the evaluation of the Pacesetter math course was largely descriptive, and focused on 45 teachers and a sub-sample of 24 teachers. The Pacesetter curriculum appeared to be more effective for some students than others. Those who did the best generally liked math and had done well in math courses in the past. Because of the work load and the novel ways for students to work, Pacesetter may be more appropriate for honors students. Overall, results of the study suggest that evaluation of curricular reform may be quite problematic, due to a lack of appropriate assessments, difficulties in assessing student growth, contextual factors, and the constraints in soliciting participation from teachers and schools. (Contains seven tables and six references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 419 012

Constraints and Limitations in Evaluating Math Curricular Reform Efforts: Pacesetter Math Case Study

Wayne J. Camara

The College Board

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Wayne Camara

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

- ☑ This document has been reproduced as received from the person or organization originating it.
- ☐ Minor changes have been made to improve reproduction quality.
- • Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Draft 3/24/98

TM028312

Paper presented at the Annual Meeting of the American Educational Research Association,

San Diego, April, 1998

BEST COPY AVAILABLE

Introduction

This symposium discusses many of the challenges and strategies in evaluating curricular reforms in secondary schools. My presentation focuses on the College Board's Pacesetter Math course, a fourth-level course that was entering its third year in 1995-96 when we undertook this pilot evaluation effort. However, much of this presentation would apply to an evaluation of any curriculum reform effort.

Pacesetter courses are offered in math, English, and Spanish and were designed by the College Board to establish high standards for all students (College Board, 1996, p. 1). The Pacesetter math course is intended to serve as an alternative to more traditional pre-calculus courses, and is designed for a broad range of students with varying interests, career intentions and mathematics preparation. The course includes a number of instructional modules and embedded assessments that emphasize three dimensions in math: (1) Knowledge, (2) Applications and Modeling, and (3) Math Communication. The culminating assessment – comprised of some multiple choice items and several extended response tasks – is completed over two course periods and is a standard part of the course. New Pacesetter teachers are required to complete a six-day intensive staff development training and many elect to attend a mid-year training institute and refresher institutions in following years (College Board, 1994).

The 1995-96 evaluation of Pacesetter math was intended to address a number of outstanding questions that the College Board and participating districts had about the program.

1. What type of students are completing the course?

BEST COPY AVAILABLE

2. Does Pacesetter "provide added value" to students?

(a) Do they have higher achievement in math as a result of the course?

(b) Do their attitudes toward the relevance of math change as a result of the course?

(c) What additional changes in students might be associated with the course and instruction?

(d) How do student-level, teacher-level, and instructional- (or pedagogical-) level differences affect these outcomes?

These questions are never as simple as they appear in any evaluation effort and certainly were not easy to address in a program entering its third year and implemented in 46 districts, 130 schools, and involving 170 teachers. As you may realize, the main issue in evaluating any program is to evaluate its effects against some baseline. That is, if you are interested in students' achievement, you may compare achievement at the end of the intervention – in this case the Pacesetter course – and compare that to achievement of a comparable group of students in non-Pacesetter courses. A second possibility is to measure student growth or change from before to after the intervention – a traditional pre- and post-comparison.

Now, why did we not apply either of these tried and true evaluation designs? Practical constraints. First, identifying an appropriate control group is difficult, but especially challenging when you employ multiple and intensive data requests throughout an entire year. In this case, to address the above questions, we realized that we would need a substantial amount of data from students and teachers both as they entered the course and on completion. Specifically, we would seek:

- demographic data on student-level and teacher-level differences
- data on student and teacher experiences and attitudes at the beginning and end of the course
- multiple measures of student achievement data at the beginning and end of the course
- data that could capture instruction and pedagogical practices used in the classroom to determine the extent teachers were following the Pacesetter model

One major obstacle we encountered was finding appropriate assessments for evaluating the achievement of the Pacesetter dimensions of learning. Pacesetter teachers, and curriculum and test development staff strongly advised us that traditional objective assessments of math achievement would be inappropriate for assessing students' abilities in: (1) Knowledge, (2) Applications and Modeling, and (3) Math Communication. They felt strongly that appropriate assessments would need to mirror the Pacesetter framework, be primarily performance-based, permit both collaborative group and individual tasks and preparation, have obvious applications to meaningful real world issues, and require the use of graphing calculators. After an exhaustive search of all College Board and ETS assessments, as well as external assessments, math content experts concluded that the Pacesetter culminating assessment was the only appropriate assessment available. Using a control group in such a design was problematic because these students would not be familiar with the framework and instructional emphases, and would clearly not perform as well on the course's culminating assessment as students completing the Pacesetter course. In addition, the culminating assessment included about 4-5 class periods of preparation and 2-3 periods for the assessment – a substantial burden for schools and teachers. Similarly,

there were no adequate pre- and post-measures to gauge student performance or student growth on two of the three dimensions.

With no available control group and no pre/post design, it is difficult to place results of the Pacesetter math course in a larger perspective. Results from the evaluation, therefore, are more descriptive, and correlation results capture only limited perspectives of the Pacesetter course and student achievement than would be desired in a full evaluation.

I will briefly summarize results from our evaluation, but these will be addressed in more detail in a symposium on Friday morning (Scheuneman and Camara, 1998; Turner, 1998; and Wilder and Cline, 1998). Some of the characteristics of math reforms efforts, including the belief systems of curriculum specialists and educators who develop and implement these reform efforts, often seriously constrain essential evaluative efforts needed to support the reforms. In the case of Pacesetter math, there was a strong belief that assessing these students on measures that did not correspond closely to the framework (in content and format) would disadvantage the students and lack validity about student achievement and the efficacy of the course. Traditional pre-calculus assessments were dismissed as not aligned to the framework (perhaps emphasizing math knowledge, but really not tapping applications/modeling and communications). Similarly, content experts felt that several external assessments claiming to reflect NCTM standards and other aspects of Pacesetter were simply not appropriate because they too lacked a close alignment with both the framework and the collaborative nature of Pacesetter. However, results from our evaluation illustrated that the three Pacesetter dimensions were highly correlated, both on the culminating assessment (r =) and teacher ratings of student competence (r =). While

BEST COPY AVAILABLE

educators are convinced that these three dimensions can be distinguished from a pedagogical perspective, assessments and teacher evaluations illustrate that such distinctions in student performance are not as easy to come by.

A side note about this study and the design. Last year the College Board embarked on an evaluation of the Pacesetter English course. We went to extraordinary efforts to include control groups of students from schools in the same district where Pacesetter English was conducted (using incentives for teachers and students) and employed released items from NAEP, essays from the Advanced Placement examination, and administered a partial form of the Pacesetter culminating assessment to both groups to provide more comparable data between groups as well as over time (Harris and Smith, 1998).

Methods

As noted above, the evaluation attempted to describe Pacesetter students and to examine student outcomes resulting from the course (e.g., achievement, attitudes, intentions) and contextual factors that may influence the outcomes. There levels of contextual factors were identified: (1) student background and math preparation (e.g., courses completed, grades, ethnicity, and gender); (2) teacher background and experience (e.g., years of experience in teaching and in Pacesetter courses); and (3) implementation (e.g., pedagogy, use of textbooks).

Wilder (1998) explains the Pacesetter population, sampling strategies, and methodologies employed in this study in much more detail. I will only briefly address general outcomes of the evaluation. First, a 50 percent sample of 80 teachers was selected for a broad survey of student

and teacher backgrounds, attitudes, and instructional practices. Of these 80 teachers, 45 teachers provided the necessary data for the analyses. Surveys were administered at the beginning of the course and at the end of the course. A more intensive sample of 24 teachers was selected from the original 80 teachers; they were asked to:

- complete the fall and spring teacher and student surveys
- administer a traditional Algebra test to students in the fall
- participate in site visits and interviews conducted in late winter
- administer a traditional math test in the spring (testlet from the SAT II math level IIC test)
- complete ratings of each student on the three Pacesetter dimensions
- provide final course grades for students
- administer the Pacesetter culminating assessment

The fall examination was the College Board's Intermediate Algebra Skills test, part of the Multiple Assessment Programs and Services™ (MAPS™). This examination contains 30 four-option items covering areas of algebra, geometry, equations and inequalities, and applications. The test used in the spring was developed for this study with 25 five-option items selected from the SAT II math level IIC test. These items cover algebra; geometry, including coordinate geometry; trigonometry; and functions.

Results

The results for the fall and spring achievement tests for the 502 students with data on both exams are shown in Table 1. The correlation between the two tests was .60.

Place Table 1 about here

The fall achievement test appears to be of appropriate difficulty for the group. The mean score is 20.1, roughly .4 standard deviation units above a middle difficulty reference value of 18.75.[1] The spring achievement test was rather difficult for the group, however. Here the mean score was 7.3, more than one standard deviation below a middle difficulty score of 12.5 (Scheuneman and Camara, 1998).

Nearly all survey respondents reported taking geometry and second year algebra, and about half had completed trigonometry (232 or 52 percent). Far fewer students completed probability or statistics (50 or 11 percent) and calculus (21 / 5 percent). About nine percent of students reported that they did not take any course beyond second year algebra. Students were classified according to the highest course taken for comparisons of fall and spring achievement test scores in Table 2.

Place Table 2 about here

---

[1] Since the fall test has no correction for guessing, the middle difficulty reference value is higher than half the number of items, taking into account the possibility that some items would be answered correctly by chance.

BEST COPY AVAILABLE

In addition to the spring achievement test, 474 students (94 percent) also had scores on the culminating assessment for math knowledge, applications/modeling, and math communication, and teacher ratings in math reasoning, applications/modeling, and math communications. Course grades for Pacesetter math were available for 437 students (87 percent). Intercorrelations and means and standard deviations for the culminating assessments, teacher ratings, and course grades are shown in Table 3.

Place Table 3 about here

The within method correlations are high, ranging from .62 to .88 for the three culminating assessment measures and from .72 to .91 for the three ratings. The ratings are also more highly correlated with course grades than are either the traditional achievement measures or the culminating assessments, suggesting that some "halo" effect may be in operation (Scheuneman and Camara, 1998). The correlations between the two application modeling measures and the two math communications measures are substantially lower. The traditional spring achievement test and the math knowledge culminating assessment score are correlated .67, suggesting more commonality of measurement than was the case with the ratings.

Finally, regression analyses using the fall achievement test score as a covariate were conducted on clusters of variables as follows:

1. ***Personal student variables***. These included age, gender, racial/ethnic background, and father and mother education. Language background was considered, but found to be

difficult to interpret due to the varied backgrounds of those reporting that English was not their best language.[2]

2. ***General academic background variables***. These included self-reported grades, year of graduation, and number of courses taken in a number of different curricular areas.
3. ***Math background variables***. These included self-reported grades in math classes, the math courses taken, and whether algebra was taken before the ninth grade.
4. ***Attitude variables***. These included scores on attitude toward math from the fall and spring surveys, change in attitude from fall to spring, and attitude toward the Pacesetter course from the spring survey.
5. ***Classroom variables***. These were the frequencies of activities in the classroom as reported by the students on the spring survey.
6. ***Student behavior variables***. These included amount of time spent on homework and days of class missed reported on the spring survey.

Because a large number of variables had to be considered, Scheuneman and Camara (1998) used a series of analyses to successively reduce the number of variables to be considered in the regression models. The following sets of analyses were performed for each of the seven outcome measures, the traditional spring achievement test, the three culminating assessments, and the three teacher ratings:

---

[2] When crossed with racial/ethnic background, the 50 students who reported a best language other than English included 6 Asian students, 6 Hispanic, 3 Black, 27 White, 3 other, and 5 who did not identify themselves with regard to ethnic background.

BEST COPY AVAILABLE

1. Separate regression analyses were performed for each of the six clusters. Those variables from each cluster that contributed to prediction of one of the dependent variables were retained in a separate data set, which was used for all subsequent analyses.

2. Variables from all clusters were then placed together in a series of step-wise regressions to further reduce the number of variables available. Because the number of students included in the analyses varied as a result of the variables included in the step-wise list, the results sometimes changed even though some of these variables were never included in the model. Only those variables entering the step-wise results for at least one of the outcome measures were retained, and the analyses were repeated to obtain the best set of predictors for each outcome.

3. Analyses were repeated, including only those variables in the best set of predictors for a particular outcome measure. All variables in the final models had significant regression weights at the .05 level and all F statistics for the final models were significant well beyond the .001 level.

4. In order to evaluate the relative importance of the variables to the prediction model, each variable appearing in the final model was successively eliminated and then replaced to determine the loss in total variance accounted for when that variable was removed. The results for the traditional spring achievement test and the culminating assessments are given in Table 4 and for the teacher ratings in Table 5.

Place Tables 4-5 about here

As expected, the fall achievement test measure of mathematics achievement was an important predictor of later achievement for all seven outcome measures. Previous grades in mathematics were nearly as important or more important that the fall achievement test scores in

predicting teacher ratings. Grades in math or overall grades appeared in all seven of the final models, and attitude toward math or toward the Pacesetter math classes figured in six models. Racial/ethnic background appeared in five of the final models (Whites achieving higher scores on the three culminating assessments and the rating for applications/modeling, and Blacks achieving lower scores on the traditional spring achievement test even in these analyses). Students completing more vocational courses and students with more days of school missed also performed more poorly on 3-4 models. The results for the traditional spring achievement test and the culminating math knowledge measure were quite similar, suggesting they were measuring similar constructs although educators had initially warned that traditional assessments would be a poor proxy for the Pacesetter framework. The regression models for both measures had nine significant predictors that together accounted for about half the variance in each. Clearly, the most important aspect of math knowledge as measured by either test is previous math achievement as measured by the fall achievement test scores, previous math grades, taking algebra prior to the ninth grade, and taking calculus (Scheuneman and Camara, 1998). The other culminating assessments for applications/modeling and math communication had results that were quite similar. The four most important of the five predictors were the same – the fall test score, overall grades, attitude toward the Pacesetter class, and White racial/ethnic group. The amount of variance accounted for by the predictors was generally less than for the two math knowledge measures, about 35 percent for applications/modeling and 27 percent for math communication.

Results from the fall and spring surveys were less clear. Student attitudes toward math and Pacesetter were less positive in the spring than the fall, although few differences were statistically significant. However, students were significantly more likely to agree with the

statements "If I had a choice I would not study more math" ($t = 6.4$, $p<.001$) and "would not take another course like this one" ($t = 4.4$, $p<.001$), and less likely to agree with the statement "I like math" at the end of the course than in the fall. Approximately 85 percent of all students were in their senior year of high school, which may partially explain their more negative attitudes and ambivalence toward math as the year progressed (Wilder and Cline, 1998).

Students who performed best in the Pacesetter course were those who were best prepared in terms of achievement (fall achievement test, courses taken, math grades), who liked math generally as measured by the fall survey, and liked Pacesetter as measured by their spring survey. In general students with positive attitudes performed better on all three dimensions of the course. Positive attitudes among these students were further associated with the teacher's implementation of the curriculum, particularly problem solving in groups. This finding is of some interest as previous research has not shown a consistent relationship of math achievement with group problem solving.

Conclusions

The Pacesetter curriculum appears to be more effective for some students than for others. Those who do best are those who generally like math and have done well in math courses in the past. Second year algebra only does not appear to be adequate preparation for the course contents, and students with this level of preparation generally did poorly. Teachers have even suggested that this courses, designed as a fourth-level math course for all students, may be more appropriate for honors students (Turner, 1998) because of the work load and novel ways for students to work.

Some of the defining features of Pacesetter – emphasis on collaborative learning and problem solving, discussion of alternative solutions to problems, talking and writing about math, and the applied nature of learning and assessments (College Board, 1994) – may have been sufficiently removed from the more traditional experiences of students to have created some anxiety and negative attitudes among students when introduced at their senior year. Interviews and site visits suggested that students were concerned about how Pacesetter would prepare them for more traditional courses in math in their future and noted the difficulty in working in unfamiliar ways (Wilder and Cline, 1998).

Overall, the results of this study suggest that evaluation of curricular reform efforts can be quite problematic. The lack of appropriate assessments that closely reflect the content and pedagogy emphasized in the curriculum, the difficulties in assessing student growth, contextual factors (e.g., student background, instructional practices), and the constraints in soliciting participation from teachers and schools that are often over-evaluated and over-tested (both schools participating in the reform and control sites) can impact the validity and utility of evaluation efforts. Despite these limitations, we can make some tentative recommendations about Pacesetter math, such as: (1) the curriculum and instructional practices may need to become more standardized in how they are presented; (2) better student preparation should be required for the course; and (3) many of the defining features of the Pacesetter course may be more appropriate and accepted if introduced earlier in a student's math experiences.

References

College Board (1994). Precalculus through modeling. Pacesetter teachers edition, 1, 1-223.

Harris, C.D. and Smith, D. (1998). Evaluation of Pacesetter English. Alexandria, VA, Human Resources Research Organization.

National Council of Teachers of Mathematics (1989). Curriculum and evaluation standards for school mathematics. Reston, VA. The Council.

Scheuneman, J.D. and Camara, W.C. (1998). Analysis of student achievement in the Pacesetter mathematics course. Paper presented at the Annual Meeting of the American Educational Research Association in San Diego.

Turner, H. (1998). Mathematics teaching in the 1990's: Are Pacesetter teachers teaching from the standards? Paper presented at the Annual Meeting of the American Educational Research Association in San Diego.

Wilder, G. Z. and Cline, F.A. Student Reactions to Pacesetter: data from 70 classrooms. Paper presented at the Annual Meeting of the American Educational Research Association in San Diego.

Table 1

Performance on Fall and Spring Achievement Tests
All examinees
(N=502)

	Fall Achievement Test	Spring Achievement Test
Mean Score[3]	20.1	7.3
Standard deviation	5.2	5.0
25th percentile	16.2	3.8
Median	20.5	6.2
75th percentile	23.9	10.1
Skewness	-.19	+.80
Score Range	5 - 30	-4 - 24
Mean Percent Correct[4]	67	40

[3] The scores on the fall achievement test were the number of correct responses. The spring achievement test was formula scored, that is, the raw score was the number correct minus 1/4 of the number of responses that were incorrect.

[4] Based on the number of correct items for both examinations.

Table 2

Fall and Spring Achievement Test Mean Scores for Students Reporting Different Courses taken[5]

Course	N	Fall Test	Spring Test
No course beyond 2nd year algebra	40	16.4	4.4
Precalc/3rd yr Alge.	161	19.8	6.5
Trigonometry	182	20.6	7.7
Statistics/Probability	39	20.6	8.3
Calculus	21	23.0	13.5

[5] The scores were sorted into categories by the highest course taken with courses ordered from low to high as precalculus/3rd year algebra, trigonometry, statistics/probability, and calculus.

Table 3

Intercorrelations among Fall Achievement Test and Various Pacesetter Outcome Measures (N = 474) [6]

	Culminatg Math Knowldg	Culminatg Apply Model	Culminatg Math Commun	Rating Math Reason	Rating Apply Model	Rating Math Commun	Course Grade
Fall Test	.53	.51	.43	.39	.43	.45	.35
Spring Test	.67	.59	.52	.45	.50	.53	.41
Culminating Math Knowledge		.70	.62	.50	.56	.56	.44
Apply/Model			.88	.53	.54	.55	.43
Math Communication				.49	.47	.48	.36
Rating Math Reasoning					.77	.72	.56
Apply/Model						.91	.71
Math Communication							.69
Mean	20.7	18.2	5.6	2.9	3.0	2.9	3.6
Sd	8.7	10.3	5.9	1.3	1.2	1.1	1.0

[6] The number of students with course grades was 437.

Table 4

Mean Scores on Outcome Measures for Students Taking Different Courses

	Highest Course Taken				
Measure	Algebra 2	Precalculus/ Algebra 3	Trigono-metry	Statistics Probability	Calculus
Culminating					
Apply/Model	12.7	16.0	19.3	21.3	28.3
Communication	3.4	4.6	5.9	6.4	11.4
Math Knowldg	16.3	18.9	21.6	23.5	30.0
Rating					
Apply/Model	2.3	3.0	3.0	3.2	3.7
Communication	2.1	3.0	3.0	3.0	3.6
Reasoning	2.3	2.7	3.1	3.2	3.7
Course Grade	2.9	3.6	3.7	3.6	4.0

Table 5

Regression Analysis Results
Spring Achievement Test and Culminating Assessments

Percent of Variance in Prediction Lost When Variable is Removed

	Spring Achievmt Test	Culminating Math Knowlg	Culminating Appl/Model	Culminating Math Commun
Fall Test	12.9	8.5	16.0	11.9
Math Grades	3.5	2.4		
Overall Grades			3.4	2.4
Fall Attitude	3.4	2.2		
Pace Attitude			4.7	5.2
# Tests (-)	4.5	1.4		
White		2.9	1.2	2.1
Black (-)	1.6			
Calculus	1.6	2.4	0.8	
Mother Educ		2.0		
Father Educ	1.2			
Pre-9 Algebra	1.1	1.1		
Technical Sem	1.0			
Vocatnl Sem (-)				1.0
Days missed (-)		0.7		
No. Variables	9	9	5	5
R^2	50.6	45.4	34.5	27.0
N students	298	285	332	336

Table 6

Regression Analyses (cont.)
Teacher Ratings

Percent of Variance in Prediction Lost When Variable is Removed

	Applications/ Modeling	Math Communication	Math Reasoning
Fall Test	8.1	7.6	5.6
Math Grades	15.4	6.5	4.7
Textbook	1.9	2.6	
Days Missed (-)	0.5		4.2
Trigonometry			1.6
All Grades		1.7	1.2
White	1.0		
Pacesetter Attitude	0.6	1.4	
Vocational (-)	0.8	1.2	
No. Variables	7	6	5
R^2	48.5	51.3	31.0
N students	340	346	338

Table 7

Regression Analyses
Pacesetter Course Grade

Loss in Prediction when Variable is Removed

	All Students	Boys	Girls
Math Grades	15.8	24.6	16.5
Days Missed	4.9	9.2	
Fall Test	2.4	4.7	1.4
Gender (Girls +)	2.1	--	--
Overall Grades	1.3		3.1
History Semesters (-)	1.4	5.1	
R^2	47.0	47.0	45.7
N	300	133	224